Astronomical Data Analysis Software and Systems XVI
ASP Conference Series, Vol. 376, 2007
R. A. Shaw, F. Hill and D. J. Bell, eds.

VOSpace: a Prototype for Grid 2.0

Matthew J. Graham

California Institute of Technology, Pasadena, California, USA

Dave Morris, Guy Rixon

Institute of Astronomy, Cambridge, UK

Abstract. As Grid 1.0 was characterized by distributed computation, so Grid 2.0 will be characterized by distributed data and the infrastructure needed to support and exploit it: the emerging success of Amazon S3 is already testimony to this. VOSpace is the IVOA interface standard for accessing distributed data. Although the base definition (VOSpace 1.0) only relates to flat, unconnected data stores, subsequent versions will add additional layers of functionality. In this paper, we consider how incorporating popular web concepts such as folksonomies (tagging), social networking, and data-spaces could lead to a much richer data environment than provided by a traditional collection of networked data stores.

1. Introduction

Grid 2.0 is a term coined earlier this year (Gibbs 2006) to describe "a new world of distributed ubiquitous virtual computing, networking and storage in the enterprise that will allow a whole raft of new rich services." Although this might be construed as just yet more jumping onto the 2.0 bandwagon, it does actually have some substance: whereas Grid 1.0 is all about shared compute cycles, Grid 2.0 focuses on data—data sharing, and the computing, networking and storage needed to access and use it. That this is not just vaporware is proven by the take-up of services such as Amazon S3[1] which are being developed to under-pin this *web of data*. The Grid 2.0 paradigm also blurs the distinction between producer and consumer since it enables participation, and this leads to a much richer user experience.

Within the VO, the IVOA Grid and Web Services Working group is developing an interface to distributed data called *VOSpace*, and this paper considers what could be achieved if we were to apply Grid 2.0 concepts to it.

[1] http://aws.amazon.com/s3

2. VOSpace

VOSpace[2] is the IVOA specification for a distributed storage mechanism that provides a uniform interface to existing or new underlying storage implementations (the Facade pattern). It distinguishes between structured and unstructured data and offers specific functionality for these (e.g. data format conversions, such as VOTable to comma-separated variable), that can be applied to structured data. Within VOSpace, each data object is represented as a node with a unique URI identifier. Each node has a map of key-value properties and a list of data views (formats) that the node can accept or provide. The space specifies and negotiates data transfer protocols and employs WS-Security, and instance-specific access policies, for authentication and authorization, respectively. There are methods to access service metadata, create and manipulate nodes and their metadata, and transfer data.

The current version of the specification (v1.0) describes a flat, unconnected data store (something akin to an anonymous FTP directory) and there are reference implementations at Caltech, Cambridge, and ESO. The next version will add support for containers (directories), links (for federating VOSpaces), and searching, but if this is done properly then VOSpace could be much more than just a collection of networked data stores.

3. Properties Revisited

A Grid 2.0 concept that has been popularized by sites such as del.icio.us and Flickr is the *folksonomy*, where users create and share their own private custom tags on data objects. In VOSpace, a data object can have an arbitrary number of properties attached to it, where a *property* is just a string value with a URI identifier. The URI for standard properties, such as content size, will refer to something that is resolvable in a VO registry, but obviously one does not want to register custom properties. However, a user could store a data object (create a node) in VOSpace describing the custom property, and then use the URI identifier of this node as the URI identifier for the property. For example, the document *xray-properties.xml* could describe various X-ray related tags such as whether an object is an X-ray candidate or not. This document could then be stored in VOSpace with an identifier such as `vos://some.namespace/xray-properties.xml` and the candidacy property could then be referred to as `vos://some.namespace/xray-properties.xml#candidate`.

Now interesting nodes (data objects) will have lots of custom properties (tags) attached to them, and this could be an inconvenience when dealing with the node, e.g., when doing a full listing. There is also the issue of access permissions when attaching a tag to the node, since it might not be world-taggable. If, instead, a user created a link to the node then they could apply all their tags to the link. This is already the model that is used by existing services that support tagging.

---

[2]http://www.ivoa.net/twiki/bin/view/IVOA/VOSpaceHome

4. Searching

Traditionally if someone wanted to make a large data set available to the world, they would give a copy to a large data center who would publish and curate it on their behalf. However, services such as VOSpace are enabling an age of distributed personal data publication where everyone can expose their own data, and datasets can be constructed from holdings in distributed resources. The management of data within these data-spaces, or webs of loosely connected data sources, is critically dependent on the metadata that are exposed, and on searching and querying it.

The Grid 2.0 way of expressing metadata is as a Resource Description Framework[3] (RDF) triple consisting of a subject-predicate-object expression. Using this model, a custom tag in VOSpace could be represented as: node URI–property URI–property value. It would then be possible to explore the data by searching the tags using SPARQL[4], the W3C query language for RDF. A sample query might be to be find the first 25 data objects in the Caltech and Cambridge VOSpaces that have been tagged as being detected in the radio (`vos://some.namespace/radio-properties.xml#candidate = "yes"`) and in the X-ray (`vos://some.namespace/xray-properties.xml#candidate ="yes"`):

```
PREFIX xprop: <vos://some.namespace/xray-properties.xml#>
PREFIX rprop: <vos://some.namespace/radio-properties.xml#>
SELECT DISTINCT ?ivoid
FROM <http://nvo.caltech.edu/vospace-1.1/find>
FROM <http://ast.cam.ac.uk/vospace-1.1/find>
WHERE {?source xprop:candidate "yes" ;
               rprop:image "yes" ;
               vos:ivoid ?ivoid . }
LIMIT 25
```

Search results could also be ranked based on their *interestingness*[5]: this is not just how many tags an object has associated with it, but is an attempt to quantify what makes a particular data object interesting. Every user action on a piece of data expresses something about its semantic content or its relationship to other data. By analyzing this user activity (what tags does the data have, what queries are matching the data, what else are the data being collated with, who is examining/accessing the data and when) some kind of measure can be determined and associated. Interestingness is also clearly something that changes with time.

---

[3]http://www.w3.org/RDF/

[4]http://www.w3.org/TR/rdf-sparql-query/

[5]http://www.flickr.com/explore/interesting/

5. Social Networking

Ultimately data exploration can lead to new collaborations as new data sets are discovered. For example, a user might search for radio data on objects for which they already have X-ray observations, but they might require assistance if they are not particularly skilled in radio astronomy; and so they may form a working partnership with the creator of the radio data. Of course, such associations need not be so haphazard: FOAF[6] is an RDF-based mechanism for expressing personal profile-type information. Using this, common research interests could be expressed, and collaborations formed, by intelligent agents on the user's behalf. Search results might also identify possible collaborations by recommending other data products: for example, 57% of users who were interested in this data set were also interested in this other data set.

Moving the collaboration layer down to the machine level might also serve to redefine data access policies. The two extreme scenarios are a *communist* model where all data is open (such a model is proposed by LSST) and a *capitalist* approach where data is traded on markets (e.g. a UKIDSS data set is worth so much Pan-Starrs data).

6. Conclusion

Although some of the ideas in this paper might seem fanciful, there are clear benefits to incorporating Grid 2.0 concepts into VOSpace. Exposing semantic information about data facilitates data exploration and discovery and can make this a dynamic process. This, in turn, can lead to new collaborations, better data access and a richer data experience.

References

Gibbs, T. 2006, Grid Today, (5:16), available at:
http://www.gridtoday.com/grid/629435.html

[6]http://xmlns.com/foaf/0.1/